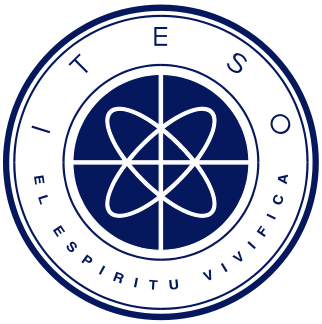

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

DEPARTMENT OF ELECTRONICS, SYSTEMS, AND INFORMATICS

COMPUTING SYSTEMS ENGINEERING

MACHINE LEARNING COURSE

TEACHER: EDNA L. GUEVARA RIVERA

BIRD STRIKE FATALITY PREDICTION ON AIRPLANE CRASHES

PROJECT DEFINITION

PRESENTED BY:

MARCO RICARDO CORDERO HERNÁNDEZ, 727272

CARLOS EDUARDO RODRÍGUEZ CASTRO, 727366

SEPTEMBER 20TH, 2022

AUTUMN, 2022

TLAQUEPAQUE, MÉXICO

MACHINE LEARNING COURSE

Index

Introduction

It is often said that one is more likely to die in a car crash than in an airplane accident. This isn't an exaggeration, even being backed up by the United State National Safety Council [1]. This fact it's usually accompanied by the contrast of the highly likeliness of having an automobile accident in the way to an airport rather than in the plane itself. These aspects are backed up by the common knowledge of what is required in order to get a pilot's license versus the minimum aspects needed for a driver's license. Setting aside the economic resource needed for wings to fly, not anyone can become an airplane pilot, not even a private one, and those who do, they need to be in constant training [2]. For this sole reason, the probability of being in an aerial incident, fatal or not, it's very low. But, what about when there is indeed an accident? It can't be denied that human factor plays a big role in the final outcome of an aerial sinister, whether it's from land by air traffic controllers or by pilots stunned by unusual conditions [3]. Although it might seem contradictory to the first lines of this paragraph, the reality is that even by staying extremely calm, the most prepared and experienced cabin crew can't deal with a motor failure or complete loss in its entirety. This can be aided by analytics.

Machine learning (from now on referred as ML), as revised by Brown [4], may be seen as "the capability of a machine to imitate *intelligent human behavior*". Given this short but meaningful definition, the problem that this project will try to tackle can be seen as this: humans cannot think fast enough in a matter of life and death, whereas computers could certainly do.

By giving a proof by counterexample, Gupta [5] details two scenarios in which ML should be avoided thoughtfully: fairly ease or complexity lacking problems, and lack of labeled data. To put in someone's hands the life of several people it's not something to be taken lightly. The beautiful field of applied math conjoined with computer science usage could potentially save hundreds if not thousands of lives; just by applying simple algebra concepts such as matrices and dot products [6] great things can be achieved, solutions can be made and existing methods of avoiding fatalities can be drastically improved, in this case, through the application of ML. Although this it's just the introductory part of this work, it can be assured that poorly classified data or niche information won't be a problem in the becoming development.

Furthermore, and getting into a deeper level of detail, it's almost immediately recognized that the problem found can be addressed by applying supervised learning algorithms; as defined by Richards & Jia [7], these classifying algorithms make quantitative analysis over a dataset to decide whether an entry or set of entries correspond to some type of classification. This type of classification it's called like so because in order for it to work, desired outputs have to be given.

With the previous being said, it is not without reminding that ML it's just as strongest as its weakest link. ML it's a powerful tool, but it won't do miracles. In any case, the following sections will explore specific portions of the whole project.

Problem to solve

Ever since it happened, the US Airways flight 1549, or the "Miracle on the Hudson" as its often referred to, has become the flagship of aircraft incidents that turned out well in terms of fatalities. [8] [9] On January 15, 2009, said flight suffered a *bird strike* which led to a successful water landing, in which only injured passengers were reported, this meaning that no deaths were suffered on the incident. This is extremely rare, as the odds of surviving a plane crash versus those of an aquatic emergency landing are completely different [10]. At the moment of the incident, Chesley Sullenberger, the pilot that made the maneuvers for the successful landing, had over 40 years of experience or *training*, key factor in the fortunate outcome of the situation. With this in mind, does it really take a flight veteran to make or predict a favorable result in terms of lives lost?

Perhaps it might seem harmless at first glance, but when organic material such as birds' corpses get stuck into complex and carefully engineered machinery such as airplane turbines or helicopter rotors, disastrous events take place. The broken components of these aircrafts can be easily diagnosed with modern on-board systems, a detail of vastly interest, because with this piece of information, severity can be predicted ipso facto.

As a form of summarization, this project seeks to predict the fatality of a bird crash incident over type of aircraft, having such outcomes as *fatal* (0) and *non-fatal* (1).

Data collection

A dataset containing 25558 registers and 26 features has been retrieved from a data science platform [11].

The description of said set states that the values contained within the dataset comes directly from the Federal Aviation Administration (FAA), who provided the number and details of incidents where birds have struck a plane over a period of ten years, this being from 2000 to 2011 (two years after the Hudson incident).

With aid from pandas (a popular python data analysis library [12]), a quick analysis was made in order to determine the absence of values, which, in this case, was indeed found.

```
RangeIndex: 25558 entries, 0 to 25557
Data columns (total 26 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   record_id                              25558 non-null  int64
 1   aircraft_type                          25558 non-null  object
 2   airport_name                           25558 non-null  object
 3   altitude_bin                           25558 non-null  object
 4   aircraft_make_model                    25558 non-null  object
 5   wildlife_number_struck                 25558 non-null  object
 6   wildlife_number_struck_actual          25558 non-null  int64
 7   effect_impact_to_flight                25558 non-null  object
 8   flightdate                             25558 non-null  object
 9   effect_indicated_damage                25558 non-null  object
 10  aircraft_number_of_engines             25291 non-null  object
 11  aircraft_airline_operator              25558 non-null  object
 12  origin_state                           25109 non-null  object
 13  when_phase_of_flight                   25558 non-null  object
 14  conditions_precipitation               25558 non-null  object
 15  remains_of_wildlife_collected          25558 non-null  bool
 16  remains_of_wildlife_sent_to_smithsonian  25558 non-null  bool
 17  remarks                                20787 non-null  object
 18  wildlife_size                          25558 non-null  object
 19  conditions_sky                         25558 non-null  object
 20  wildlife_species                       25558 non-null  object
 21  pilot_warned_of_birds_or_wildlife      25558 non-null  bool
 22  cost_total                             25558 non-null  int64
 23  feet_above_ground                      25558 non-null  int64
 24  number_of_people_injured               25558 non-null  int64
 25  is_aircraft_large                      25558 non-null  bool
dtypes: bool(4), int64(5), object(17)
```

Fig. 1 Overview of dataset and analysis of values

The previous results show that the features number of engines, origin state and remarks not only have null values, but also that they all are objects, most likely strings.

Learning type to use

A supervised categorical algorithm has been chosen for this type of problem because there needs to be determined if a flight accident will or will not be fatal and our output will always be between "Fatal" and "Not fatal", ideally. In this case, the type of algorithm that'll be developed is categorical because it needs to classify all of the results under one of these two categories that are set.

One of the greatest advantages that this type of algorithm will bring to the project is that the result will be easily readable and no further processing is needed to extract real value from the output. Despite this, the algorithm has one disadvantage, debugging a categorical algorithm can be harder since there cannot be explicitly seen that an issue exists. The issue can only be detected when tests are made from the predicted results; since there is no complete control over specific operations the algorithm is doing, the debugging process can be quite time consuming.

When doing the comparison between the two main algorithm contenders (regression and categorical), discovers were made: even though regression can be considered a more precise algorithm, it lacks the output simplicity that the categorical algorithm is known for. All of the research points to use the categorical algorithm to predict whether a flight accident is fatal or not, the pros outweigh the cons for this specific application.

Conclusions and pending work

Through this document, the first stone has been set for the incoming project that would encapsulate the knowledge gathered along the course for which this text has been written.

The most prevalent piece of work that needs to be made it's the transformation of the dataset as demonstrated in previous sections. Text or string fields were present, and, although this could be seem as problematic, the reality is that this information needs to undergo over a transformation and cleaning process in which these categorical data would be transformed into numerical values. The scope of said process goes beyond the purpose of this introductory file, but it certainly will appear in the future over reported advances.

References

[1] National Safety Council, «Odds of Dying,» NSC Injury Facts, 2020. [Online]. Available: https://injuryfacts.nsc.org/all-injuries/preventable-death-overview/odds-of-dying/. [Last access: 14 September 2022].

[2] K. Hoke, «AeroSavvy,» AeroSavvy , 25 April 2018. [Online]. Available: https://aerosavvy.com/recurrent-training/. [Last access: 14 September 2022].

[3] Clifford Law Offices PC, «The National Law Review,» The National Law Review, 8 December 2020. [Online]. Available: https://www.natlawreview.com/article/most-common-causes-aviation-accidents. [Last access: 2022 September 14].

[4] S. Brown, «MIT Sloan School of Management,» MIT, 21 April 2021. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. [Laste access: 14 September 2022].

[5] S. Gupta, «Springboard,» Springboard, 25 September 2020. [Online]. Available: https://www.springboard.com/blog/data-science/when-not-to-use-ml/. [Last access: 14 September 2022].

[6] Uniqtech, «Medium,» Data Science Bootcamp, 22 December 2018. [Online]. Available: https://medium.com/data-science-bootcamp/understand-dot-products-matrix-multiplications-usage-in-deep-learning-in-minutes-beginner-95edf2e66155. [Last access: 2022 September 2022].

[7] J. A. Richards y X. Jia, «Supervised Classification Techniques,» de *Remote Sensing Digital Image Analysis*, Berlin, Springe, 1999, pp. 181-222.

[8] C. Eastwood, Dirección, *Sully.* [Movie]. United States: Flashlight Films, 2016.

[9] S. Lanfermeijer, «Tailstrike,» Tailstrike Consultancy, [Online]. Available: 2022. [Last access: 15 September 2022].

[10] D. Null, «The Guardian,» 2011. [Online]. Available: https://www.theguardian.com/notesandqueries/query/0,5753,-10081,00.html. [Last access: 2015 September 2022].

[11] J. Shih, «data.world,» 2016. [Online]. Available: https://data.world/shihzy/2000-2011-birds-strikes-planes. [Laset access: 15 September 2022].

[12] pandas, «pandas documentation,» 12 September 2022. [Online]. Available: https://pandas.pydata.org/docs/. [Last access: 15 September 2022].